7Д

УДК 004.62

Г.В. Дорохина
Институт проблем искусственного интеллекта МОН Украины и НАН Украины
г. Донецк, Украина
sgv@iai.donetsk.ua

Сравнение затрат памяти для метода деревьев цифрового поиска и его усовершенствования

В статье выполнен сравнительный анализ затрат памяти для организации поиска строковых величин методом деревьев цифрового поиска и его усовершенствования. Разработана методика теоретической оценки затрат памяти для обоих методов. Выполнено сравнение реальных данных с расчетными оценками при заданном количестве узлов древовидной структуры.

Введение

Одну из наиболее высоких скоростей поиска обеспечивают деревья цифрового поиска (ЦП). Данный метод используется редко ввиду больших затрат памяти, т.к. он предполагает хранение древовидной структуры и самих данных в специально выделенной области памяти. В связи с этим *актуальным* является анализ способов организации данных, позволяющих выполнять поиск данных с той же скоростью, что и деревья ЦП, и сократить затраты памяти по сравнению с этим методом.

В работе [1] предложено усовершенствование метода деревьев ЦП, обеспечивающее хранение строковых величин в древовидной структуре и скоростной поиск данных в ней. Оно характеризуется такой же скоростью поиска данных, как у метода деревьев ЦП, и за счет хранения данных в древовидной структуре может давать значительный выигрыш в затратах памяти.

Целью данной работы является сравнительный анализ затрат памяти при использовании деревьев ЦП и усовершенствованных деревьев ЦП. Поставленная цель может быть достигнута путём решения следующих *задач*: формальное описание затрат памяти на хранение данных и деревьев ЦП; формальное описание затрат памяти на хранение данных методом усовершенствованных деревьев ЦП; сравнительный анализ затрат памяти.

Затраты памяти при поиске строковых величин методом деревьев цифрового поиска

Как отмечено ранее, метод деревьев ЦП предполагает хранение данных, а также древовидной структуры, обеспечивающей их скоростной поиск. Рассмотрим задачу использования деревьев цифрового поиска для задачи поиска элементов словаря строк – множества строк без повторений. Для хранения самих данных можем использовать массив строк фиксированной длины или строк переменной длины.

Пусть максимальная длина строки в словаре равна h, а словарь содержит z строк. При его хранении с помощью строк фиксированной длины необходимо

$$P\,(G) = (h + 1) \cdot z \cdot Sc \qquad (1)$$

байт. Здесь Sc – количество байт, отводимых под хранение одного символа.

Затраты памяти массива строк переменной длины для хранения того же словаря при известных величинах z_i – количество в словаре G строк длины i вычислим согласно:

$$P(G) = \sum_{k=1}^{h} z_i \cdot (2 + 4 + k \cdot Sc). \quad (2)$$

При этом 2 байта отводится для хранения реальной длины строки, а 4 байта – для указания ссылки на строку.

Быстрый поиск данных методом деревьев ЦП обеспечивает древовидная структура. От содержимого словаря зависит $v = \sum_{i=1}^{h} v_i$ – количество узлов в ней. Здесь v_i – количество в древовидной структуре узлов высоты i. Определим границы данной величины, исходя из данных: t – размер алфавита символов, составляющих строки словаря; z – количество строк в словаре G; z_i – количество в словаре G строк длины i; $l = [\log_t z]$ – целая часть от $\log_t z$.

$$\sup(v_i) = \begin{cases} t^i \mid i \le l \\ \sum_{j=i}^{h} z_j \mid i > l \end{cases}, \quad (3)$$

$$\inf(v_i) = \max\left(z_i, \left]\frac{v_{i+1}}{t}\right[\right). \quad (4)$$

Здесь $\left]\frac{v_{i+1}}{t}\right[$ – операция округления вверх результатов деления $\frac{v_{i+1}}{t}$.

Исходя из верхней, нижней границ v_i и вышевведенного значения l, величина v удовлетворяет:

$$\inf(v) = \sum_{i=1}^{i=h} \inf(v_i), \ \sup(v) = \sum_{i=1}^{l} t^i + \sum_{i=l+1}^{h} (i - l) \cdot z_i. \quad (5)$$

Здесь $l = [\log_t z]$.

Каждый из множества узлов O дерева цифрового поиска может быть описан вектором $o_j = (o_{j1}, o_{j2}, o_{j3})$, где o_{j1} – символ алфавита; o_{j2} – целое число, $o_{j2} = 0$ для узла, в котором не оканчивается ни одна строка, и $o_{j2} \in [1; z]$ для узла, в котором оканчивается строка с номером o_{j2} в словаре G; o_{j3} – множество номеров узлов, дочерних по отношению к j-му узлу.

Объем памяти, необходимый для хранения узлов высоты i представим в виде суммы объема памяти, необходимого для хранения элементов o_{j1}, o_{j2} этих узлов, и суммарного объема памяти, необходимого для хранения $\bigcup_j o_{j3}$, где j принадлежит множеству узлов высоты i. Причем

$$\left|\bigcup_j o_{j3}\right| = v_{i+1}. \quad (6)$$

Обозначив количество байт, отводимое для хранения целого числа, через Si, получим выражение для оценки затрат памяти для хранения множества узлов высоты i:

$$v_i \cdot (Si + Sc) + v_{i+1} \cdot Si. \quad (7)$$

Тогда затраты памяти для хранения всех узлов дерева ЦП:

$$P(O) = \sum_{i=0}^{h} \left(v_i \cdot (Sc + Si) + v_{i+1} \cdot Si\right).$$

Разбив эту сумму на две, получим

$$P(O) = (Sc + Si) \cdot \sum_{i=0}^{h} v_i + Si \cdot \sum_{i=0}^{h} v_{i+1} = (Sc + Si) \cdot \sum_{i=0}^{h} v_i + Si \cdot \sum_{i=0}^{h} v_{i+1} .$$

Для любого непустого словаря $v_0 = 1$, а $v_{h+1} = 0$. Перепишем выражения

$$\sum_{i=0}^{h} v_i = 1 + \sum_{i=1}^{h} v_i ,$$

$$\sum_{i=0}^{h} v_{i+1} = \sum_{i=1}^{h+1} v_i = \sum_{i=1}^{h} r_i + 0 = \sum_{i=1}^{h} v_i .$$

Получим

$$P(O) = Sc + Si + (Sc + 2 \cdot Si) \cdot \sum_{i=1}^{h} v_i . \qquad (8)$$

Подставив в (8) выражения (3) и (4), получим оценку затрат памяти для хранения дерева ЦП:

$$\sup(P(O)) = Sc + Si + (Sc + 2 \cdot Si) \cdot \left(\sum_{i=1}^{l} t^i + \sum_{i=l+1}^{h} (i - l) \cdot z_i \right), \qquad (9)$$

$$\inf(P(O)) = Sc + Si + (Sc + 2 \cdot Si) \cdot \sum_{i=1}^{h} \max\left(z_i , \left] \frac{r_{i+1}}{t} \right[\right), \qquad (10)$$

где $l = [\log_t z]$, где $r_{h+1} = 0$,] [в – операция округления до большего целого числа.

Использование деревьев ЦП предполагает хранение древовидной структуры и самих данных. Поэтому совокупные затраты памяти при использовании метода деревьев ЦП $P(G,O)$ определяются суммой

$$P(G,O) = P(G) + P(O) . \qquad (11)$$

Затраты памяти при поиске строковых величин методом усовершенствованных деревьев цифрового поиска

Усовершенствование [1] метода деревьев ЦП состоит во введении в узел древовидной структуры дополнительного элемента данных – номера родительской вершины, а также в использовании ссылок (массива номеров узлов) на узлы-окончания строк. Это усовершенствование позволяет хранить данные в самой древовидной структуре. То есть, с одной стороны, увеличивается размер узла и появляется массив ссылок на узлы-окончания строк, с другой стороны, исчезает необходимость в хранении массива данных наряду с древовидной структурой.

Каждый из множества узлов O' дерева цифрового поиска может быть описан вектором $o'_j=(o_{j1}, o_{j2}, o_{j3}, o_{j4})$, где элемент o_{j4} – номер родительского узла по отношению к данному. Массив ссылок (номеров) на узлы-окончания строк обозначим через X. Его размер равен количеству строк в словаре z. Причём

$$\forall x_i \in X \, \exists \, o'_k : k = x_i , o_{k4} = i .$$

Количество узлов в древовидной структуре, построенной по одному словарю, совпадает для дерева ЦП и для усовершенствованного дерева ЦП.

Обозначим через $P(O', X)$ затраты памяти для хранения усовершенствованного дерева ЦП.

$$P(O', X) = P(O') + P(X) . \qquad (12)$$

$$P(X) = z \cdot Si$$

Найдем объем памяти, необходимый для хранения множества O'. В каждый элемент o_j' множества O' в сравнении с элементом o_j множества O дополнительно введен

элемент o_{j4}, являющийся целым числом. Значит, размер каждого элемента o_j' превышает размер o_j на размер целого числа Si. По аналогии с деревом ЦП:

$$P(O') = Sc + 2 \cdot Si + (Sc + 3 \cdot Si) \cdot \sum_{i=1}^{h} v_i \,, \tag{13}$$

$$\sup(P(O')) = Sc + 2 \cdot Si + (Sc + 3 \cdot Si) \cdot \left(\sum_{i=1}^{l} t^i + \sum_{i=l+1}^{h} (i-l) \cdot z_i \right), \tag{14}$$

$$\inf(P(O')) = Sc + 2 \cdot Si + (Sc + 3 \cdot Si) \cdot \sum_{i=1}^{h} \max\left(z_i, \left] \frac{r_{i+1}}{t} \right[\right). \tag{15}$$

Сравнение затрат памяти метода деревьев цифрового поиска и метода усовершенствованных деревьев цифрового поиска

Проанализируем изменение затрат памяти при фиксированных размере алфавита и объеме словаря, но различных значениях h. Пусть $z = 1000000$, $t = 33$. В формулах определения затрат памяти фигурирует величина z_i. Допустим, эта величина определяется по формуле

$$z_i = \min\left(t^i, \frac{z - \sum_{j=1}^{i-1} z_i}{h - i + 1} \right), i \in \overline{1, h-1}, z_h = z - \sum_{j=1}^{h-1} z_i \,, \tag{16}$$

что соответствует наиболее равномерному распределению величин z_i.

Вычислим выигрыш в затратах памяти метода усовершенствованных деревьев ЦП по формуле

$$f = \frac{P(G,O) - P(O',X)}{P(G,O)} \cdot 100\% \,. \tag{17}$$

При оценке затрат памяти будем рассматривать три случая:

– количество узлов в древовидной структуре максимально $v = \sup(v)$;
– количество узлов в древовидной структуре минимально $v = \inf(v)$;
– количество узлов в древовидной структуре является средним между минимальным и максимальным $v = \frac{\inf(v) + \sup(v)}{2}$.

На рис. 1 отражены зависимости относительного выигрыша в затратах памяти (17) метода усовершенствованных деревьев ЦП от максимальной длины строки словаря (h по оси аргумента), количестве в словаре слов длины i, определяемом по (16) для минимального, максимального и среднего количества узлов.

Приведенные данные относительно выигрыша (проигрыша) в затратах памяти усовершенствованного метода деревьев ЦП имеют слишком большой разброс. Это делает затруднительным его оценку. Кроме того, выбранное распределение длин строк в словаре не соответствует реальному распределению строк в словарях, элементами которых являются слова языка.

Приведем пример расчетных и реальных относительных затрат памяти для строк фиксированной и переменной длины. В качестве словарей были использованы база словоформ русского языка, содержащая 1 779 843 уникальных строк, и база начальных форм слов русского языка, содержащая 110 746 уникальных строк. База словоформ в отличие от базы начальных форм содержит большее количество уникальных строк. В то же время она содержит большее количество строк, начальные символы которых совпадают и отлич-

ны только несколько последних символов. То есть одна и та же ветвь древовидной структуры позволяет хранить начальные символы большего количества слов. Распределение длин уникальных строк в этих базах приведено на рис. 2. Видим кардинальное отличие распределений длин слов реальных словарей от равномерного распределения (16), которое мы использовали при подсчете затрат памяти для хранения различных представлений словарей. Поскольку рассмотренные реальные словари содержат меньше слов большей длины, то затраты на их хранение должны быть меньше приведенных ранее расчетных.

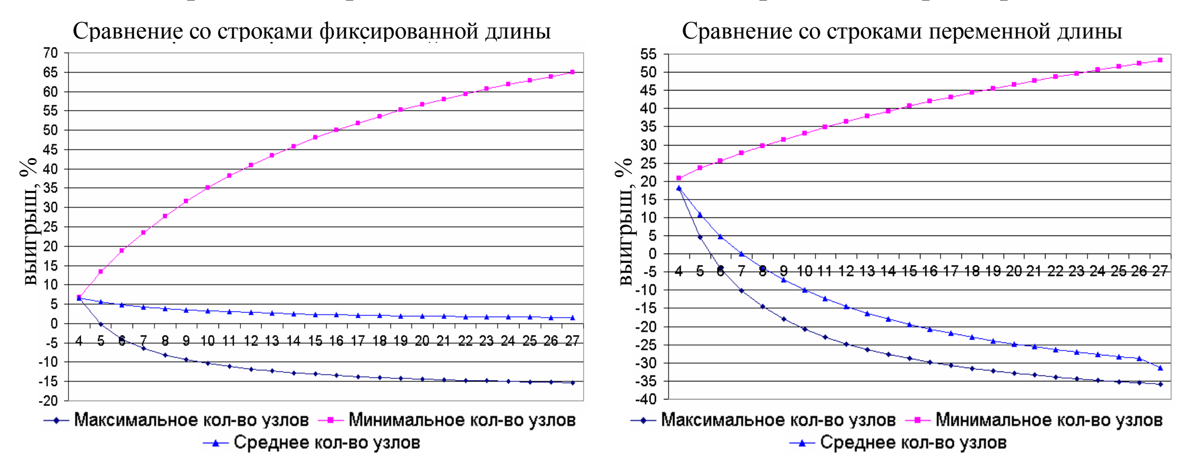


Рисунок 1 – Относительный выигрыш в затратах памяти
метода усовершенствованных деревьев ЦП

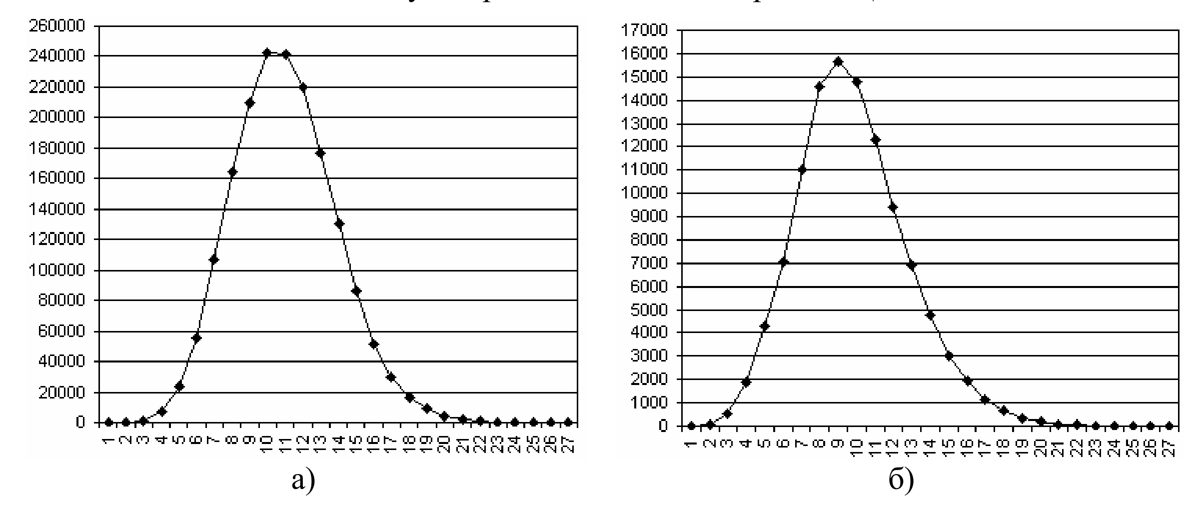


Рисунок 2 – Распределение длин строк в реальных словарях: а) в базе словоформ
русского языка; б) в базе начальных форм слов русского языка

Сопоставим полученные нами реальные данные выигрыша в затратах памяти разработанного метода с расчетными данными при минимальном, максимальном и среднем числе узлов в древовидной структуре (табл. 1).

Видим, что при хранении базы словоформ – словаря большего размера с большим количеством строк, отличающихся друг от друга несколькими последними символами, – выигрыш от использования разработанного метода приближается к расчетному выигрышу при минимальном количестве вершин древовидной структуры. При хранении же базы начальных форм оцениваемый выигрыш приближается к расчетному выигрышу при среднем количестве вершин. Отметим также, что метод усовершенствованных деревьев ЦП может проигрывать в затратах памяти базовому методу, если словарь строк базового метода хранится в виде строк переменной длины.

Таблица 1 – Выигрыш в затратах памяти на словарях слов русского языка

Словарь	Длина строк	Реальные данные	Расчетные данные при заданном количестве узлов древовидной структуры		
			Max	Min	Ave
База словоформ	фиксир.	43,10	-6,86	54,05	15,94
	перемен.	22,60	-20,57	34,55	-7,77
База начальных форм	фиксир.	17,30	-4,80	54,05	18,29
	перемен.	-4,32	-20,89	31,36	-8,50

Заметим также, что мы оценили выигрыш разработанного метода при хранении словарей, состоящих из уникальных строк. В базе начальных форм общее число строк – 111 133, а мы рассматривали словарь из 110 746 уникальных строк этой базы. По базе словоформ мы сформировали словарь, содержащий 1 779 843 уникальных строк, тогда как общее число строк в этой базе – 3 167 168.

Выводы

В работе рассмотрена проблема оценки затрат памяти усовершенствованных деревьев ЦП. Предложена методика теоретической оценки относительного выигрыша в затратах памяти этого метода в сравнении с базовым при известных характеристиках словаря: размер алфавита символов, образующих строки словаря; максимальная длина строки словаря; количество в словаре строк заданной длины. Полученная методика позволяет определить целесообразность использования для словаря с заданными параметрами метода усовершенствованных деревьев ЦП. Сравнение реальных данных согласуется с теоретическими оценками при заданном количестве узлов древовидной структуры. Анализ показывает эффективность применения усовершенствованного метода деревьев ЦП для хранения и поиска строковых величин, различающихся последними символами.

Литература

1. Патент України № 78806. Пристрій для збереження і пошуку рядкових величин та спосіб збереження і пошуку рядкових величин / Дорохіна Г.В.; патентовласник: Інститут проблем штучного інтелекту МОН України і НАН України // Промислова власність ; опубл. 25.04.2007, Бюл. № 5.

Г.В. Дорохіна

Порівняння витрат пам’яті для методу дерев цифрового пошуку та його удосконалення
У статті проведено порівняльний аналіз витрат пам’яті для організації пошуку рядкових величин методом дерев цифрового пошуку та його удосконалення. Розроблено методику теоретичної оцінки витрат пам’яті для обох методів. Порівняно реальні дані з оцінками, що обчислено при заданій кількості вузлів деревоподібної структури.

G.V. Dorokhina

Memory Expenses Comparison for the Method of Digital Search Tree and Its Improvement
The paper is devoted to the problem of memory expenses for the method of digital search tree and its improvement. The method for theoretical estimation of memory expenses of these structures is proposed. Comparison of the real memory expenses and calculated estimations are made.

Статья поступила в редакцию 24.06.2009.